Learner construction of corpora for general English in Taiwan

Smith, S.

PROOF COVER SHEET

Author(s): Simon Smith
Article title: Learner construction of corpora for general English in Taiwan

Article no: NCAL_A_557024

Enclosures: 1)Query sheet
2)Article proofs

Dear Author,

1. Please check these proofs carefully. It is the responsibility of the corresponding author to check these and approve or amend them. A second proof is not normally provided. Taylor & Francis cannot be held responsible for uncorrected errors, even if introduced during the production process. Once your corrections have been added to the article, it will be considered ready for publication.

For detailed guidance on how to check your proofs, please see
http://journalauthors.tandf.co.uk/production/checkingproofs.asp

2. Please review the table of contributors below and confirm that the first and last names are structured correctly and that the authors are listed in the correct order of contribution. This check is to ensure that your name will appear correctly online and when the article is indexed.

Sequence	Prefix	Given name(s)	Surname	Suffix
1		Simon	Smith	

Queries are marked in the margins of the proofs. Unless advised otherwise, submit all corrections and answers to the queries using the CATS online correction form, and then press the "Submit All Corrections" button.

AUTHOR QUERIES

General query: You have warranted that you have secured the necessary written permission from the appropriate copyright owner for the reproduction of any text, illustration, or other material in your article. (Please see http://journalauthors.tandf.co.uk/preparation/permission.asp.) Please check that any required acknowledgements have been included to reflect this.

AQ1 **Please provide department/division name for the affiliation.**

AQ2 **Some authors have experienced problems when their home email addresses have been published. Please can you confirm that you are comfortable with this address being made public?**

AQ3 **Please spell out IT, ELT, ICT, BNC, ESL, CALL, POS in full at first mention.**

AQ4 **Please confirm the addition of the heading ‘‘Introduction’’ per journal style.**

AQ5 **Kindly note that the figure citations and the figures have been renumbered to make them sequential.**

AQ6 **Please provide details for ‘‘Notes on contributors’’ section.**

AQ7 **Has reference ‘‘Boulton, in press’’ been published yet? If so, please give details for reference section.**

AQ8 **Figures 1, 2, 4, 5, and 6 looks more like a table. Hence, kindly suggest whether this can be set as table.**

Computer Assisted Language Learning
Vol. 00, No. 0, Month 2011, 1–26

Routledge
Taylor & Francis Group

Learner construction of corpora for general English in Taiwan

Simon Smith*

NCCU, Taipei, Taiwan

This exploratory study describes a framework for data-driven learning (DDL), in General (nonmajor) English university classes, in which learners *construct* linguistic corpora instead of merely *consulting* them. Prior related work has addressed the needs of language specialists, in particular trainee translators who are learning how to compile glossaries, rather than nonmajor students of English. It is argued in this article that the process of creating a corpus inculcates a sense of ownership in the learner and therefore has a motivational impetus. This is especially true, it is claimed here, when the topic of the corpus is of personal interest to the learner, or coincides with their major field of study. Learners may pursue language study for only a short period of their university career but once the corpus is constructed, some students may be sufficiently motivated to consult it and add to it when needed. Moreover, the process of compiling the corpus may lead to the acquisition of not only language but also useful transferable skills, including IT and problem-solving competencies. This study presents some of the motivational issues surrounding DDL in Asia and suggests corpus construction as a solution. Previous research on corpus construction by learners is reviewed. In the experiment which forms the core of this study, 90 freshmen compiled and analyzed corpora as part of a General English course in Taiwan. Of these, 19 students completed final projects based on corpora they had compiled. Their findings – and reactions to the use of corpora compilation as a language learning tool – are reported in a qualitative data analysis.

Keywords: ELT; Taiwan; DDL; corpus construction

Introduction

In recent years, Western ELT scholarship has paid considerable attention to student-centered and computer-based learning, focusing in particular on the pedagogical use of linguistic corpora. The use of linguistic corpora in language learning often takes the form of concordance analysis by students or data-driven learning (DDL). Johns (1991) likens the language learner (on the DDL model) to a researcher, analyzing target language data and becoming familiar with the language through the regularities and consistencies encountered.

As Johns explained, DDL attempts to impart linguistic knowledge by making available samples of authentic language, from corpora, and inviting language learners to tease out or discover usage patterns for themselves, effectively acting as

*Email: smithsgj@gmail.com

ISSN 0958-8221 print/ISSN 1744-3210 online

DOI: 10.1080/09588221.2011.557024
http://www.informaworld.com

researchers rather than mere language students. The present article describes a refinement of DDL, whereby the learners were given tools to build and analyze their own corpora. The work is novel in two respects. While much has been written on the creation of corpora by learners, any previous work on their creation by nonspecialists seems to have gone unreported (although Kennedy & Miceli, 2001, describe and evaluate the *consultation* of corpora by nonlanguage majors). Furthermore, it is one of only a handful of qualitative DDL analyzes based on Taiwan data.

In what follows, some of the challenges for English learning in Asia are reviewed, with respect in particular to learner motivation, and how this has affected the adoption of corpus based approaches here. I address motivational and cultural issues that affect learning choices, and make a case for corpus construction as a pedagogical task that holds considerable promise, which may yield better results for some learners than other DDL task types. The importance of *ownership* of the deliverable, a corpus, is discussed and it is argued that corpus construction is one way to acquire useful transferable skills. There follows an extensive review of the literature on corpus construction by learners and a description of the WebBootCat (WBC) software used by the students in my class.

Then, I present some of the content and outcomes of a semester-long course featuring corpus construction by nonspecialist learners. The analysis is entirely qualitative in nature and draws on homework submitted by students, who not only reported on interesting findings from their corpora but also provided some commentary on the usefulness of homemade corpora, as compared to large reference corpora, as well as what they thought about the task itself. The article concludes with a discussion of limitations of the research and presents plans for future work.

The study attempts to answer the following research questions:

(1) Do students find corpus construction and consultation enjoyable and useful?
(2) Does their English improve as a result?
(3) Do they have a sense of ownership of the corpus?
(4) Do they consult or modify the corpus after the course ends?
(5) Do they acquire any transferable, nonlinguistic skills, such as IT, critical thinking or problem solving?

Motivational challenges in Asian classrooms

In the universities of Taiwan, and many Asian nations, those teaching English to nonmajors have to overcome certain hurdles if they are to successfully use DDL or other corpus-based techniques. The detractors of DDL, according to Johns (1991), believe only ''intelligent, sophisticated, and well-motivated'' students will benefit from DDL. While not going as far as to claim that unintelligent, unsophisticated, or poorly motivated students will get anything out of DDL, he does note that one might be surprised at the achievements of ''most students[,] given the opportunity to show what they are capable of.'' (p. 12)

Obviously, Taiwanese undergraduate students are no more or less intelligent across the board than their counterparts elsewhere, and it would be inappropriate to make general claims about their degree of sophistication; it is widely accepted among English teachers in Taiwan, though, that students who are not majoring in English studies are generally not well motivated to learn the language. Lai (2008) found that

university student motivation often wanes over the course of a semester, while Ho (1998) attributes low motivation (in high schools) to the fact that English is a compulsory subject which "has little to do with the daily life of ... students." Perhaps some of the blame for the lack of motivation can be apportioned to teachers: Cheng and Dörnyei (2007) find that in Taiwan "most practising teachers ... do not deem adopting interesting learning tasks a significant component of motivating learners."

In a study of two Asian students with differing needs (not from Taiwan), Turnbull and Burston (1998) highlighted the impact of both motivation and degree of specialization on the success of concordancing. One, a student of Applied Linguistics, who planned to remain in Australia after the course, adapted well to the DDL task, and successfully used concordances as a language learning tool. The other, studying Public Policy and Management, had no real need to improve his English for academic or professional life. He found it difficult to figure out what to do with corpus data and appeared to prefer a traditional approach to language learning. The author's teaching experience would place many General English students in Taiwan in the latter category: they have little immediate need for English and no plans to live in an English-speaking country.

Can DDL work in Asia?

Hadley (2002) used printed concordance materials with intermediate Japanese students. Despite his own initial reservations, he did report some success: many of his students believed that DDL helped their studies. Others, though, found the approach difficult and "incoherent."

A potential objection to DDL in Taiwan or Japan is that learning through reflection and discovery does not seem to fit too comfortably into the received model of Asian pedagogy. Most stakeholders (students, parents, university administrators, education officials, and unfortunately many teachers) have a fairly traditional understanding of the learning process, where essentially the teacher delivers content and the students somehow absorb it (or the students practice production, and the teacher corrects them where necessary). Sun and Wang (2003) bemoan their finding that Taiwanese teachers find inductive learning "too time-consuming." The situation prevails despite the best efforts of Taiwan universities, both private and public, to encourage and reward student-centered learning and learner autonomy (Savignon & Wang, 2003). Cheng and Dörnyei (2007, p. 168), indeed, found that the motivational strategy of "promoting learner autonomy" is hardly used in Taiwanese EFL contexts.

A second difficulty for DDL in Taiwan lies in the type of study exercises often associated with the approach. Given the foregoing paragraphs, the reader might be surprised to learn that exercises which would be acceptable to Western learners might well be too dull for Taiwanese consumption. Let us explain. In the years spent studying English before university, most students will not usually have learned communicative skills, and will have had only rare opportunities to speak English. Vocabulary and grammar are taught by memorization and exposure to pattern sentences, followed by exercises, typically consisting of gap-fill questions or error correction.

Now, university freshmen are aware that there are problems with the high school approach; although they are often not very motivated to acquire communicative

skills in English, they do expect to be taught in a way that is markedly different from their high school experience. The last thing one would expect them to want is more gap-fills and error correction exercises. Ironically, DDL solutions do often take these forms (Johns, 1991, 1997); indeed, Thurstun and Candlin (1998) and Boulton (2008) commend gap-fills precisely because of their familiarity to students.

The sorts of gap-fill exercise set in high school textbooks are not merely inauthentic; they are often downright implausible and sometimes even unacceptable to native speakers. The fact that gapped concordance lines to be completed *are* on the whole authentic, though, is lost on many students. De-contextualized, the lines are often difficult to make sense of, even if students are encouraged to work in sentence rather than keyword in context (KWIC) mode, such that sentences, rather than sentence fragments, can be viewed in a concordance. In the absence of motivation to learn, corpus investigation is unlikely to lead to the sense of achievement experienced by some learners at serendipitous linguistic discovery (Bernardini, 1997; Cheng, Warren, & Xu, 2003).

Rising to the motivational challenges

The motivational picture is not all bleak, however. Students do enjoy certain types of creative and productive tasks, especially when they feel they have ownership of the outcome. Students asked to make a short film featuring the members of their group, for example, will generally have a better time and learn more than if they had been asked to simply present a skit in class; students prefer to keep a regular blog, which can be read and responded to by teacher and peers, than turn in traditional writing assignments. There seem to be two main reasons for the preference. First, a permanent record is kept: there is a tangible (or at least digital) object which can be viewed and reviewed and shown to others; and, as noted, the object is then *owned* by its creators. This sense of ownership probably does not extend to (say) an ordinary writing task, graded and returned by the teacher.

Second, the students have to deploy nonlinguistic skills to complete the task: they need to learn how operate the camera to best effect and how to save, store, and play the resulting digital file. In the case of blogging, they need to know how to navigate the blog interface, how to add pictures, and change fonts, and formatting. Useful transferable skills are often acquired, although the extent to which this happens will of course depend on the student's background. Jackson (1997) lists a number of such skills that his Computer Aided Text Analysis students acquired: project management, problem solving, and report writing, as well as computer skills. In fact, some of his students signed up for the course in order to try to overcome their own "computer phobia" (p. 237). Other writers have studied the interplay between the learning of language and of ICT skills, in particular Cheng et al. (2003) who describe a corpus linguistics course intended to bridge the gap between ICT and discourse analysis, "two formerly discrete subjects" (p. 177). Boulton (in press) found that his students learnt a lot about ICT while working on their corpus-based projects, in particular more effective use of the Internet and advanced features of Microsoft Office products. Of course, ICT is not for everyone but it is safe to say that young adult learners, in Taiwan at least, enjoy the challenge of using technology for assignments.

Using corpus creation software presents a similar challenge as far as the use of technology is concerned. While students may find corpus *consultation* somewhat

anodyne and dry, they might be expected to take some pleasure in creating their own domain-specific corpus, particularly if it is related to their personal interests or area of study. They will certainly have ownership of the resulting corpus: they set the keywords and parameters for its construction, and they are able to use it or add to it at any time.

Braun (2005) discusses what makes a corpus "pedagogically relevant." Her remarks are made in the light of the claim by Widdowson (1978) that authentic texts – corpus materials – are only of use if they can be *authenticated* by the learner; that is to say, the learner can define or imagine a context for the text. Braun (p. 53) suggests that some of the conditions for authentication will have been satisfied if the corpus is "relevant for the needs of the target group," and gives "genre-specific corpora" as one way of meeting those needs. A corpus created by students, in the domain in which they specialize, certainly qualifies as "relevant".

Low motivation to learn English notwithstanding, one call that *will* in most cases be made upon the students' English is the need to understand textbooks and other academic writing. Although in most cases Taiwanese students only take English for the first year of their university careers, they will often need to draw upon their knowledge of the language in subsequent years. The knowledge needed will often be the terminology and usage of the academic discipline in which they happen to be majoring; the availability of a domain-specific corpus will give students access to the sublanguage in which they are interested, as well as a sense of how it differs from general English. The corpus can be extracted from current, relevant documents on the Web, and can be updated and consulted at any time.

The learner-built corpora in this study, therefore, are pedagogically relevant (because they are in the student's area of specialization) and constitute a concrete resource from which the student might be expected to derive a sense of ownership.

Prior research on corpus construction

This is by no means the first time that domain-specific corpora have been used for language learning. Aston (1995) assembled small corpora from CD-ROM collections of texts and assigned exercises on collocation and grammatical patterns. Tribble (1997) demonstrated "quick and dirty" ways to assemble 30–40 thousand word themed corpora, using the Microsoft Encarta software. Both writers make the point that small, on-topic corpora like these are potentially more useful for language learning than concordances from large generic corpora, which, while authentic and representative of the language, may overwhelm learners with their unpredictable and sometimes incomprehensible contexts.

The themed corpora reported by Tribble (1997) were built by teachers (as opposed to learners, as in the case of the present study) as part of their materials development. In later work, Aston (2002) compares the relative benefits of published corpora and those which are *homemade* by either teachers or the learners themselves. He finds in favor of a hybrid solution, where a learner extracts a domain- or usage-specific subcorpus from a published corpus, noting too that experience in corpus compilation is likely to sharpen the learner's creativity and critical awareness.

What follows is a brief review of the literature on learner-built corpora: usually action research, where the teacher has instructed students to build a corpus in the hope that they will acquire linguistic or other skills, either during the building process or in subsequent use – or preferably both.

Tyne (2009) reports on a course in which British students were asked to create spoken corpora of French. This differs somewhat from the present work, in that Tyne's students' task was to record authentic French, using native speakers. They then had to prepare a transcription. The goal was not, primarily, to improve the learners' French, although it turned out, unsurprisingly, that most students did feel their French had got better. The task was assigned as part of a sociolinguistics class, to teach the students about variation in language. Tyne emphasizes the importance of the process of corpus building, and the students' sense of "ownership" of their corpus which emerges from that process: "Le corpus apparaît ici non pas comme une simple source de données pour l'enseignant ou l'apprenant, pouvant être interrogée par un ensemble d'outils, mais comme une ressource, fabriquée par l'apprenant qui va servir dans l'apprentissage de la langue tout en nourrissant une réflexion sociolinguistique sous-jacente." [This kind of corpus is much more than just a data source for teacher or learner; it is a resource of the learner's own making, helping him or her to acquire language, while at the same time cultivating the kind of sociolinguistic reflection which underpins acquisition.] (Tyne, 2009, p. 93)

Boulton (in press) presents a corpus linguistics course, for French postgraduate students of English, which he designed and taught. His students were mature individuals with strong motivation to learn and a willingness to do so autonomously. As with the course taught by Tyne (2009), Boulton is more interested in the learning process than the final product. Students had the responsibility for "defining the question, choosing the corpus, finding the appropriate tools, using their ingenuity to overcome the problems they [would] inevitably encounter" (p. 3). In this case, the final product is a corpus-based piece of research, not a corpus per se. However, in choosing their data sources, students had the option to use published or homemade corpora. Students were also free to use both types, for comparative purposes. In all, 27 of the 30 participants took up the option to create their own corpus, either from the Web or from some other source. Of that number, 15 decided to use published corpora as well. One of the students compared simplified and unabridged versions of the same English novel, using vocabulary profiling software. Several of the projects analyzed rock music lyrics, while others examined specific linguistic points.

Seidlhofer (2002) describes the use of a collaborative learner corpus in her class of advanced and highly motivated trainee English teachers, making the students' own work the "primary objects of analysis" (p. 217). This approach, she claims, encourages the students to engage in noticing, linguistic hypothesis testing, and metalinguistic reflection. Attending to these procedures, according to Swain's (1995) *output hypothesis*, makes learners "more likely to modify their output, and do so successfully." Seidlhofer's students were asked to write a summary and a commentary on a *Time* article she supplied. The students' work was collected in electronic format, names were deleted and a representative sample distributed to all students. They then – in small groups – compiled a list of questions about the writings, which addressed aspects of lexis, grammar and style as well as content. Spurred on by the fact that they had themselves produced the writing, and wanting to know how it could be improved, the students were enthusiastic about using the corpus tools to which they had been introduced, including concordancing, to answer some of the questions.

There is a significant body of work on student-built corpora in terminological and translation studies. Because professional translators are often given assignments which lie outside their own area of expertise, it is important for them to understand

unfamiliar technical terms in the source document, as well as to be able to translate them accurately into the target language. There are specialist dictionaries and other resources available to help them but they are not always sufficiently comprehensive or up to date for the purpose. A resource which does not suffer from these two constraints is the World Wide Web; given the means to compile assignment-specific corpora from the Web, and thereby to extract relevant terminology, trainee translators would clearly be well equipped to handle assignments in unfamiliar domains.

Maia (1997) had her Portuguese-English translation students build themed corpora, for the purpose of compiling domain glossaries. Initially the students transcribed paper documents and used CD-ROMs (including Encarta), but once Web access was available to the class, that became their principal data source. Domains explored by the students included current affairs (electoral systems; war and conflict) and other topical issues such as ecology and IT. In some cases, existing glossaries found on the Web were enhanced, in others the terms were extracted from domain texts by the students.

Castagnoli (2006) also discusses the importance of the Web as a resource for translators and terminologists, containing as it does a vast number of texts on practically all subjects, and in nearly every register. Because there is no upper bound on the size of a corpus that can in principle be generated from the web, corpus users are unlikely to find that their queries return few or zero results. The web also offers the advantage that most translators – and indeed most language learners – are already familiar with it and use it regularly.

Castagnoli's terminology students, who were translators in training, used the BootCaT toolkit (Baroni & Bernardini, 2004) to generate web corpora on specific topics, and extract lists of terms which could be used to compile glossaries and term databases. The students found that a larger number of relevant terms could be extracted when the domain chosen was highly specialized. Thus, a *company law* corpus yielded a lot of vocabulary in that domain, while a *cellphone* corpus contained a higher proportion of general terms. Castagnoli ascribes this disparity to differences in the nature of websites found in the different domains: company law websites are likely to be descriptive and factual, while cellphone websites are more likely to have a commercial or persuasive purpose. By way of assessment, the students were given a technical translation task and were asked to prepare for it by building a web corpus in the relevant domain, and extracting from it a glossary of terms.

A number of studies, then, have shown how corpus creation tasks can benefit language students. Most of these studies, though, have been based on corpora assembled by teachers or students from texts. Only Castagnoli's (2006) study reports on the automatic generation of corpora – by students of translation – from the web. The present study extends that work to the needs of nonspecialist, nonmajor undergraduate students of English.

Methodology

Software used: WBC and Sketch Engine

In the course which is the focus of the present article, students used a refined version of BootCaT to create their corpora, known as WBC. The *Web* prefix denotes that the program itself runs as a web interface: unlike the original BootCaT (which does nonetheless make use of the Web!) it does not have to be installed on the user's

computer. Baroni, Kilgarriff, Pomikálek, and Rychlý (2006), in a paper which introduces WBC, again focus on the tool's utility as an aid to technical translators. Smith, Sommers, and Kilgarriff (2008) explore the use of WBC to generate vocabulary lists for English learning.

The choice of software tool was partly motivated by the fact that the students had already been using the Sketch Engine (SkE; Kilgarriff, Rychlý, Smrž, & Tugwell, 2004), the corpus query tool in which WBC is embedded. Although SkE was originally designed for use by lexicographers and corpus linguists, many of its features help in making corpus data more accessible to language learners: see, for example, Thomas (2008) for an account of SkE in use by Czech students of English. Concordances are enhanced by making available a sentence mode, as well as the traditional KWIC mode, so that more may be gathered from the context, and allowing the display of a much larger window of context than just one line. Concordance lines can also be ranked by quality: a *good* example (GDEX) sentence is defined by Kilgarriff, Husak, McAdam, Rundell, and Rychlý (2008) as one which is neither too short nor too long, which does not contain a lot of rare words or anaphors (which can sometimes only be resolved by looking outside the sentence), among other constraining parameters. As well as this fairly powerful concordancing function, SkE offers *word sketches*, which provide a one-page summary of the contextual behavior of a word, listing collocations with frequencies and salience. There are also a statistics-based thesaurus and a *sketch differences* module which highlights the collocational differences between two similar words, among others. All modules allow the user to click through to the concordancer, and view the word or collocation in context.

WBC is a convenient tool for automatic domain-specific corpus generation. Given the choice between such a web-based tool and manual assembly of a corpus from texts, all Castagnoli's (2006) students chose the BootCat option, as noted in the previous section. The nonspecialist students in this study, it was felt, would have neither the time nor the inclination to build their corpora by browsing the Web manually.

WBC is not a free tool, and a SkE software license must be purchased to use it. Although one would have to forgo the convenience of creating and analyzing corpora with one tool, teachers might prefer to download the free BootCat software (and new graphical interface) from http://bootcat.sslmit.unibo.it/.

The WBC algorithm is conceptually simple. First, a search is seeded with any number of words selected by the user. N-tuples (for example, triples or quadruples: arbitrary selections of 3 or 4) of the seed words are sent to the Yahoo search engine. The returned web pages are then used to build the corpus. A substantial amount of filtering is done to exclude web pages which do not mostly contain running text of the language in question. Measures include rejecting pages containing too many words held on a stop list, and very short and excessively large web pages: a user interface provides control over these filters. The resulting corpus may be used in two distinct ways. First, it may be inspected with the SkE concordancer and other modules (word sketch etc.) described above, to learn about a particular word and its collocations. The student could then make a comparison with the distribution of the word in a general corpus, such as the BNC.

The second possibility is that the user can generate term lists from WBC: to do this, all words in the corpus are automatically counted and their frequencies are compared with their frequencies in a reference corpus. Words whose frequencies are

significantly higher in the created WBC corpus than the reference corpus are assumed to have a strong association with the domain implicit in the original seed words (Baroni et al., 2006). In this study, learners were assigned both SkE and term list related tasks.

The default parameters of WBC are adequate for most purposes, and the designers recommend that they should be left alone by nonexpert users. However, the user does have the option to adjust the number of seed word tuples, the number and size of web pages used in corpus construction, and various other parameters. Students selecting the WBC final project were invited to experiment with different parameter settings, and some, though by no means all, enjoyed and learned from this ''software tweaking'' experience, as will be explained in the *Results*.

SkE and WBC are complex and sophisticated, and as with any professional software, a user learning curve is expected. However, most of the functions are immediately usable by language learners, and of those students choosing the WBC project none reported have any difficulty using it. For those with lesser language skills, a Chinese version of the SkE user interface has since the study been made available.

Participants

In all, 90 students at a public university in Taiwan took part in the study. The students, virtually all freshmen, were enrolled in one of three compulsory General English classes taught by the author, a native English speaker, in the Spring semester of 2008/2009. Each class included students majoring in different disciplines within the same broad area of study (social sciences, humanities, or commerce). They completed two preparatory corpus building tasks, and received basic training in corpus study in the form of mini-lectures. Toward the end of the semester, they were allowed to choose one of three final projects, contributing 25% toward the course grade. A total of 19 students chose the corpus building option, 22 students an option on language learning websites, and the majority, 49 students, a project on the use of concordancing and other features of the SkE. It is not clear whether there was any correlation between choice of project and academic specialism; in each of the three classes, the rank order of popularity of the three options was, in any event, the same.

In Project 2, students were invited to compare and contrast a number of language learning websites and platforms, some of which had been used in class over the semester, while others were to be sought out by the student. The sites were of different types, including interactive CALL sites mainly targeting listening, for example, or vocabulary. The list also featured online dictionaries and thesauruses, large general purpose ESL sites such the BBC's and British Council's, as well as a selection of corpus sites including the BNC, and sites using corpora for ESL, such as Lextutor. Discussion of the relative usefulness of CALL and traditional language learning was also expected.

In Project 3, students gave a detailed account, with screenshots and examples, of the SkE (but not WBC), paying particular attention to one of its principal modules (concordancing, word sketches, thesaurus, and sketch differences, as they wished). They also commented on the utility of corpora as language learning tools.

It was suggested once or twice, in class, that students might want to explain what led them to make their choice of project. Of the students choosing the corpus building project, Project 1, eight students gave reasons for their choice. four students

claimed that the project was of interest because it was relevant to their own field of study. Two students expressed a desire to build their own corpus (as opposed to consulting a public corpus); these were two of the three students who indicated (as shown in Figure 6 below) that they would return to WBC and use their corpora again after the course ended. One student asserted, rather unadventurously, that the project was "easier" than the other options; perhaps not surprisingly, she was one of the two students who indicated explicitly that she would not be returning to corpus work.

The instructions for the WBC task were longer than the other two projects. Although the reason for this was pointed out at the beginning of the instructions (because the basic steps for compiling a corpus were repeated there, to help students; see Appendix) it is possible that the apparent task length could have put some students off.

Of those choosing the CALL option, the five students who commented on their reasons said essentially that they wished to examine fun, nontraditional ways to learn English. It is perhaps surprising that more students did not choose this option; many CALL web applications are great fun, and all students seemed to enjoy the CALL laboratory sessions held over the semester. In the event, almost as many students chose the apparently relatively dry corpus construction as the superficially more exciting CALL option.

Nine of those choosing Project 3, the SkE module topic, responded overwhelmingly that the fact that SkE had been used in class over the year, and they were already familiar with it, motivated their choice.

Learner tasks

WBC and SkE instruction was delivered by means of mini-lectures (10–15 minutes at a time, of a two-hour General English class). Students were given an introduction to corpora and concordancing, with some examples taken from Chinese corpora in an effort to make the material more accessible. The value of corpora as a source of authentic English was emphasized, as was the importance of learning from context and collocation, as opposed to memorizing English vocabulary and Chinese translations. To this end, the students were asked to explore the meaning and usage of unfamiliar words in their regular reading assignments, by studying common collocations in SkE word sketches. For this purpose, large corpora such as the British National Corpus and the web corpus ukWaC (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008) were used.

In addition to the mini-lectures and exposure to general corpora, two preparatory corpus building tasks were set. Interested students could then choose a more demanding, longer corpus building and analysis option as their final project. In the event, 19 of the 90 students took up this option. The sequence of teaching and learning procedures are set out in Figure 1.

Preliminary workshop

A two-hour workshop was given later in the semester where students worked individually on a number of SkE tasks, including making concordances and word sketches. One of the tasks was to create a corpus with WBC. The instructions were as in Figure 2.

Step	What	When	Notes
1.	Five 10-15 minute mini-lectures on corpora and concordancing	Weeks 2-6 of the course.	One conducted during each 2 hour English class
2.	General exposure to corpora	Throughout the course	Students encouraged to use concordancing during class preparation and other reading.
3.	Preliminary WebBootCat workshop	Week 8	Students practise using WBC, and create a small, on-topic web corpus.
4.	Corpus comparison task	Week 10	Students create a new on-topic corpus. Compare output (word sketches) from it with output from a large general corpus, such as BNC.
5.	Final project (optional, as students were given two other choices of project)	Completed by Week 18 (final week)	Students compile a new corpus based on their own subject specialism. They answer questions about the corpus, and demonstrate understanding of corpora and WBC itself.

Figure 1. Teaching and learning steps.

Build a corpus on a topic of your choice, using WBC:

1. Type in about three seed words.
2. Make sure the corpus is tagged.
3. Make a wordlist for your corpus. Is the vocabulary related to the topic you wanted?
4. Have a look at the corpus using Word Sketch or concordance.

Figure 2. WebBootCat workshop student instructions.

It is fairly clear from the WBC input form how to enter seed words, opt for a POS-tagged corpus and make a wordlist; Figure 3 shows the program interface. Students who were unsure were invited to ask during the workshop. The students' choice of seed words, however, did not always seem terribly well motivated; the

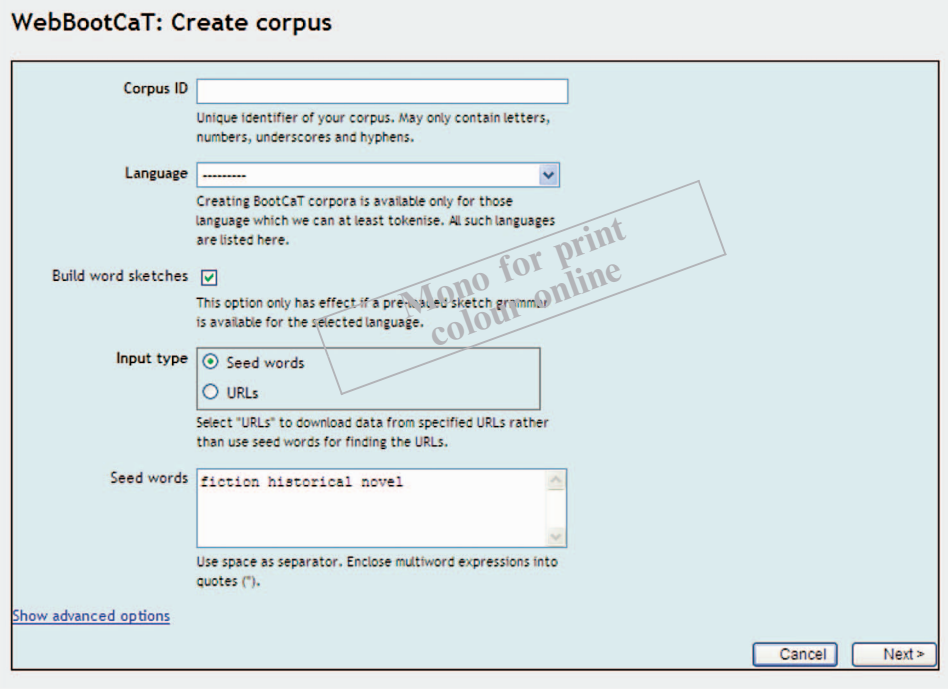

Figure 3. WebBootCat corpus construction interface.

notion of "choosing a topic" may have been unfamiliar to some students, to judge from some of the selections made. One student chose the seed words *hug, hold*, and *press*, another *bad, excellent*, and *great*. From these selections, it is not obvious what topics they had in mind, or what sort of texts they expected the search to yield. In other cases, though, students chose the sorts of seed words that one might conceivably pick when conducting an ordinary web search on a given topic: *Japan, comic, character* for texts on manga (the student probably did not use the seed word *manga* itself because he was not aware that it is used in English); *fiction, historical, novel* for a corpus on historical novels. It would probably not have occurred to the student to use the multiword term "*historical-novel*" (although WBC does in fact allow for this, as can be seen from Figure 3).

Leaving it up to students to choose their own topic led to some novel and sometimes bizarre selections: one student named her corpus "Haha" and used seed words *funny, giggle, laugh, smile*. Other corpora built included "cake," "chocolate," "ghosts," and the somewhat unsettling "killing."

Corpus comparison task

Two weeks later in the course, after feedback on the first task had been provided, the students were asked to make a comparison of the corpora they had created with a general reference corpus. The purpose was to contrast the distribution of on-topic and off-topic terms in the two types of corpus. The instructions in Figure 4 were issued, and the students were told to assume that they related to a "Police and Crime" corpus.

1. *Choose two words:*

   One on-topic (e.g. police)

   One not on-topic (e.g. car)

2. *Make 4 word sketches*

   2 from ukWaC

   2 from your large WBC corpus

3. *Comment on (1 paragraph)*

   what kind of collocations you expected to find in the word sketches.

   what collocations you actually found.

Figure 4. Corpus comparison task instructions.

The predicted answer is that the distribution of the on-topic term will not differ very much from one corpus to the other, in terms of the sorts of collocations it participate in, unless not all senses of the word carry an on-topic meaning. For example, the word *charge* has a number of senses, including the intuitively salient "take money in exchange for goods or services"; in the "Police and Crime" domain, we would expect *charge* to be used more commonly with the sense "formally accuse [of a crime]." Thus, we would expect the word *police* to have two similar distributions, while the distributions of a word like *charge* would vary somewhat.

Car participates in specialized collocations in the crime domain, for example *police car* and *getaway car*, so we might expect different distributions, here, from those of the general corpus. Given a word which does not participate in crime-related collocations, we would predict a similar distribution, or, if the word is rare, it might not occur in the on-topic corpus. Some students grasped the significance of the task, and seemed to understand that a domain-specific corpus can provide useful, topic-oriented usage data. One such student, G1, commented:

> *I choose two words, "guitar" and "style". I expect to find some collocations about music. When I use ukWaC to make sketches, for "guitar", the results are definitely conform to my expect (sic). I think that is because "guitar" is musical instruments (sic), so any kind of collocations is to be closely linked with music.*
>
> *However, "style" is different. The sketch result about this word is abundant. When I use my corpus to find collocations, I can get what I want because my corpus is "rock", which is about music. But in ukWaC, there are hair "style", architect "style", clothes "style" or life "style". The range is quite big!*

Several other students were much less certain about the relative distribution in the specialized and general corpora. One student, L1, seemed to find it problematic that in a general corpus the word *size* occurred in all sorts of contexts, while in her "sports" corpus there were "surprising" concordance lines, including one referring to the "size of tennis courts."

Final project

Students were given a project worksheet, which is supplied in full as Appendix 1. They were asked to build a new corpus, this time based on their own major subject,

and to try to optimize their corpus, by adjusting some of the parameter settings, as described by Castagnoli (2006) and Baroni et al. (2006). Students were asked to bootstrap a second corpus from the key terms generated by the first (as described in Baroni & Bernardini, 2004), and finally to answer three questions (repeated for the reader's convenience in Figure 5).

Results and discussion

In their final project, some students took the opportunity to say what they felt about the task. The five research questions addressed by this study were not explicitly posed as questions to the students, because most students would simply have responded in the affirmative if that had been done. However, 17 students did make relevant comments, and these are tabulated in Appendix 2. A summary of their opinions is presented in Figure 6.

In the following qualitative analysis, interesting highlights were selected from the completed homework and project tasks, in an attempt to show whether corpus construction is of benefit to students and what its limitations are.

Chambers (2005), in her study on corpus consultation by learners, noticed a "variation in analytical ability" among students. Not everyone, she says, was able "to reflect on the nature and limitations of the corpus." It turned out, too, that some of the students in the present study had difficulty getting to grips with the notion of the corpus as a database of authentic texts, in this case harvested from the Web: students L3, H1, R1, and T2 interpreted their results in ways that belie understanding of the structure of the corpus, or the way that the search procedure

1. To make a corpus, WBC uses Yahoo! It does an Internet search for groups of words (usually 3 words) called tuples. Try to describe in more detail how the program makes a corpus.
2. To make a corpus, is it better to just use the default WBC options (for example, 10 URLs per query)? Or do you recommend using different options? Explain how you decided, giving examples.
3. Using screen shots and other kinds of examples to help you, describe how well your corpus represents the topic you chose.

Figure 5. Extract from final project instructions.

Research question	Positive opinions	Negative opinions	Opinion not given for this question
Enjoyable/useful?	6	1	10
Helped to learn English?	8	0	9
Sense of ownership?	10	0	7
Will use after course?	3	2	12
Other skills gained?	6	0	11

Figure 6. Students who addressed the research questions of the study in their project.

works. R1, for example, despairs that she has not been able to extract the "right" meaning of *engine* from her "Romance" corpus, only instances of *search engine*; but that is precisely the context in which one might expect to find the word in such a corpus. Another student wrote of his specialized corpus "Well, I suppose there won't be any result for a word which is not on the topic, but it still has a result. And I can't figure out how it does, I can't find any relationship between the link and the topic."

The response to the parameter-optimizing part of the task was mixed. A number of students admitted that they did not really understand the parameters or that they were not sufficiently advanced users of WBC to be able to manipulate the settings to good effect. Others did experiment, and some went so far as to discuss the implications of different tuple settings (the number of seed words in each request to search the Web that WBC makes to Yahoo!). Several students observed that a very low setting, say one, will return web pages that are less relevant to the topic, because the queries were based on only one term; while a high setting (n, where n is the number of seed words specified) will generate few results, since not many documents will contain all the words. In the end, most participants concluded the most satisfactory tuple size was three, some adding that this value (the default) had in any case been fixed by experts after careful experimentation. Student D1, who built a Philosophy corpus, did, however, report better results by increasing the size of tuples.

Student T1, a historian with a special interest in Ancient Greece and Rome, found that resulting corpus was "more about Greece than History." There is no shortage of travel sites on the Web: increasing the tuple size (number of words in a tuple) might have improved her results, raising the odds that each query included a historical or academic term.

After WBC has queried Yahoo! for the seed word tuples, it presents the list of websites found to the user, who can then deselect any inappropriate URLs. This is a very important function, as it allows one to refuse corpus data from a source which is in the wrong domain, or the wrong register. Several students found that this function was useful for removing commercial sites (especially lists or databases, including the Yellow Pages site).

Turning to the representativeness of the corpora, a student of Economics (E1) found that many instances of the term *economics* itself, in the general corpora, related to school or university study of his subject. His own corpus revealed many more instances of technical or professional use of the term. Later, investigating the use of the word *elasticity*, he found that the Economics corpus contained on-domain uses of the term, while the general corpus tended to refer to it as a physical property. He concludes: "Creating a specialized corpus could be useful when it comes to researching a particular subject or learning a subject in English. It is useful because of the different results which are much more relevant than searching on a much more general English corpus." Student S2 likewise concluded that the homemade corpus was significantly more useful than the general corpus, as it included authentic terms from her subject, accountancy.

E1 also discovered an unexpected distribution for the word *car*: "For the word 'car' i didnt really think I could find a collocation in the WBC corpus, however the word car seems to come up quite often in economic examples."

L2 mentioned that he had learned new, authentic terms from his history corpus: "After finishing the project, I learned some words which I've never seen in textbook,

it's a special experience for me to learn history from Internet!" W2 was especially impressed by the inclusion of multiword economics terms.

Only one student, T2, felt that her corpus was less representative than the general corpus. She is a student of Business Administration, a domain which (arguably) is very well represented on the Web generally. Student V1 mentions the trade-off between precision and coverage, noting that both specialized and general corpora have their role. On the whole, though, students confirmed the position of Tribble (1997) and Aston (1995, 2002) that more specialized corpora have greater pedagogical value.

Some students commented explicitly on the sense of ownership they derived from building the corpus and one (P1) on the enjoyment she derived from the process of building the corpus.

> **S1**: I find it is special to have your own corpus. It is unique! You can make corpuses by your interests. That can make you know words easily because words are about your own interests. You can know the specific meaning of the words in different area (*sic*) of interests.
>
> **L1**: I think the WBC is helpful to my progress in English. The words from webs are more related to our life. In addition, WBC can produce a personal corpus. We can make a corpus which is belonging to our own. It's a corpus which we really need.
>
> **P1**: I think WBC is fun sometimes and boring sometimes. When I built my corpus I think this is fun, and feel interesting. And I feel boring sometimes, because when I watching the result of word sketch I felt it comes too many collocations and I have to extract them. The process I think is boring [The student finds the construction part more interesting than the consultation].

These comments confirm the positions of Tyne (2009) and Boulton (in press) that the construction process is more important than the final product, and that learners do indeed feel that they have ownership of the corpora they construct.

From Figure 6, observe that six students stated that they had derived some nonlinguistic benefit from the corpus construction experience and eight claimed that it helped their English. Figure 7 shows the work of a student, W1, who responded that he had benefited in both respects. He built a corpus on "Stationery" using three seed words. At the point where he took this screenshot, he is preparing to make a bootstrapped corpus by checking those terms that are in the right domain. He has missed the proper name *Waterman* (a brand of fountain pen) and the term rubber stamp, but otherwise has made a good job of identifying the on-topic items. It is unlikely that he would have been previously exposed to all of (for example) *ballpoint, Paper Mate* (another pen brand) and *graphite*, as these vocabulary items are not normally included in the Taiwan high school curriculum.

In order to bootstrap his new corpus, W1 may have used SkE to find examples of usage; he may have searched for the terms on the web, perhaps returning to the sites from which his corpus data was extracted to look at product descriptions and images. He might have simply consulted a dictionary. Whatever his approach was to the task, it is clear that he acquired some language in the process. He was expected to make a decision (on topic or off topic?) on each term, and to do this it would be necessary to reflect on the meaning of the term in a fairly focused way.

From inspection of students' work at the intermediate as well as final stages, another benefit of the corpus construction process then emerges: without the need for dedicated audit trail or logging software, it is possible for teachers to make some

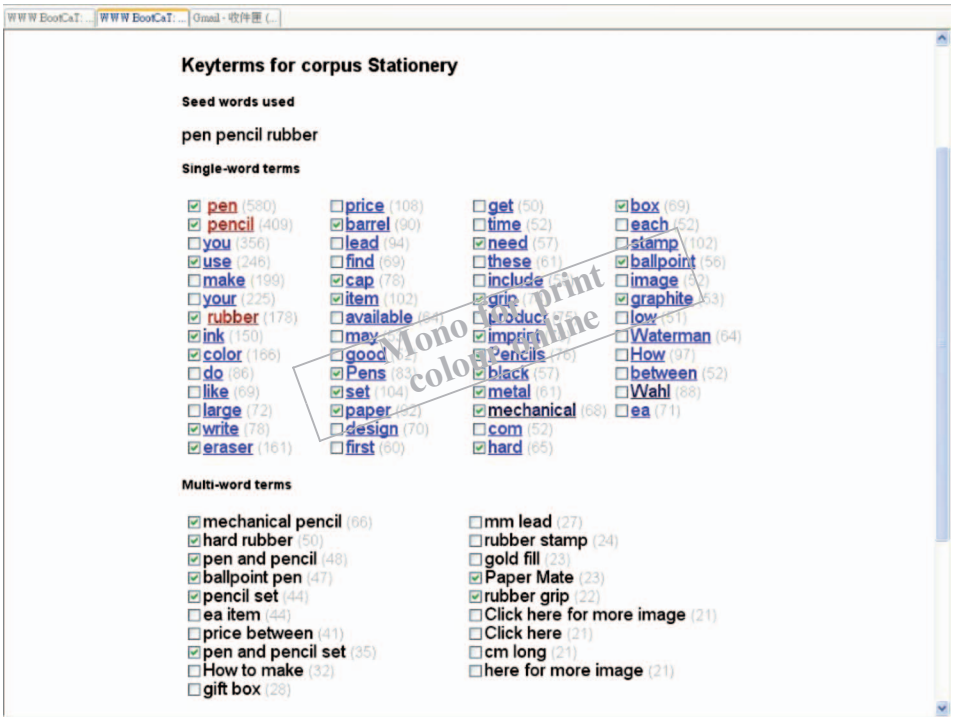

Figure 7. W1's stationery corpus.

inferences as to what learning has taken place. This knowledge could then be used for assessment or student feedback.

The extraction of spurious terms such as *Click here* (also seen in Figure 7) may well have been more of a help to the student than a hindrance. He would almost certainly have stopped to think why WBC should have picked up this "term," and that might have led him to think more about the extraction process. As mentioned above, not all students were entirely clear about the source of the keywords and corpora; possibly, the appearance of Click here served to bring the point home.

The reader will have observed that a number of students expressed some sort of disappointment or frustration with either the software or the tasks. In all, seven of the 19 projects made a negative comment of some sort. Of these, one complained that the analysis of concordances was tedious. Two others failed to understand why they would get fewer hits for a word in their homemade corpus than in a reference corpus, even though the former was supposed to be specialized (because the reference corpus is many times larger). A further two students expected to find specialized uses of terms in their corpora, but encountered mostly more general uses. Finally, two students claimed to have found bugs in the software: in one case they were correct, and the issue is being looked into by the SkE team.

Conclusions and limitations of the study

It is an important limitation of this exploratory study that the research questions were not directly addressed. It was correctly predicted, however, that a proportion of

students would address them of their own accord in the final project, given the open-ended nature of the questions. While the intentions of the author in not wishing to overtax the students with evaluation requests were good, there is no doubt that in the future a quantitative study should be conducted, in which learners are explicitly and anonymously asked their views on the research questions, or an assessment made of the learning that has taken place (by means of pre- and posttests, or similar) through the corpus construction experience.

As Figure 6 showed, though, there was some support for the research questions. It is encouraging to note the very small number of negative responses and that there were as many positive responses as this, given that no specific evaluative questions were asked of the students (other than "describe how well your corpus represents the topic you chose"; see Appendix 1). It is disappointing, though, to see that only three students mentioned that they would probably use WBC again, while a further two explicitly stated that they would not. Two of those who said they would use corpora again (as noted in *Participants* above) also said that in choosing the WBC project, they had been motivated to build their own corpus rather than consult publicly available corpora.

It was mentioned above that in total 19 of 90 students selected the corpus construction task as their final project, with the remainder choosing between two alternative options. Since all students did the earlier in-class and homework tasks, this was an informed choice, and could in principle be recruited to argue for or against students' belief in the usefulness of the task. However, since the other two tasks were also concerned with student-centered learning and computers, and discussion of those tasks is beyond the scope of this article, such an analysis was not done.

Tyne (2009) and Maia (1997) found that not everything about teaching and facilitating corpus building is predictable and straightforward, and both report that they learned a good deal along the way. The same is true of the present author – one could almost describe the process as serendipitous. In particular, although it was known in advance that more specialized domains would yield improved key term lists and richer corpora, it was not predicted just how *un*-specialized some students' domains would turn out to be, or that some would choose seed words that do not represent plausible domains. This was one factor that decided the author to constrain final project domains to students' academic major.

Opportunities for future work

It was shown above that not all this author's students were comfortable choosing the keywords which represent a domain, or even selecting a plausible domain for investigation; probably they would have benefited from some practice in formulating even relatively simple web searches. Practice with the advanced Google and Amazon.com searches, advocated by Boulton (in press), and training in more sophisticated keyword selection and query formulation would also beneficial to the students, not only as English learners but also as researchers in their home departments. Close attention to student selection of seed words and search terms will be amply repaid, and not just in terms of language learning through corpora.

WBC offers the user to evaluate and filter out certain URLs from inclusion in the corpus construction search, as noted earlier, as well as to confine corpus content

to "creative commons" pages. Although some students took advantage of this to remove listings content from the corpus, much more extensive use of the functions could be made in future research. It is not always easy for young adults to evaluate the quality or usefulness of websites, and language learners often select inappropriate sources from the Internet for their writing. The WBC URL filter could be a useful tool for training students in locating appropriate sources.

It was remarked earlier that some students may not have had a firm grasp of the nature of their corpus – what it is, and where the texts came from. While there are many benefits associated with the use of an automatic tool such WBC, if the students had been asked to garner their own texts from the Web (following Maia, 1997), or compile corpora from their own writing (Seidlhofer, 2002), there would have been no doubt about the provenance of the texts, and the students might have been able to engage better with the resulting corpora. In a future study, students could be invited to construct corpora by either (a) using WBC or the free BootCat interface or (b) individually selecting texts from the Web and university-held databases (leaving the choice up to the student, or assigning tasks to groups). It would then be possible to use WBC and SkE to examine both (a) and (b). A comparative evaluation not only of the two kinds of corpus generation, but also of the degree of insight of the students choosing each task, could then be conducted.

Only a small number of students stated that they would consult their corpora or WBC again. In fact, the participants in this study will probably not study English again during their time at university. Future studies would benefit from participants whose period of English study is longer than one year, as it would then be easier to encourage and monitor ongoing use. In this way, an account of the rather motivating *process* of corpus construction could be complemented by a more comprehensive analysis of the use to which learners put the corpus *product* than was possible in the present study.

Notes on contributors

References

Aston, G. (1995). Corpora in language pedagogy: Matching theory and practice. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson* (pp. 257–270). Oxford, UK: Oxford University Press.

Aston, G. (2002). The learner as corpus designer. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus analysis* (pp. 9–25). Amsterdam: Rodopi. Retrieved August 27, 2009, from http://www.sslmit.unibo.it/~guy/graz.htm

Baroni, M., & Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the Web. In *Proceedings of 4th International Conference on Language Resources and Evaluation*, Lisbon, Portugal, 1313–1316. Retrieved November 8, 2010, from http://people.sslmit.unibo.it/~baroni/publications/lrec2004/bootcat_lrec_2004.pdf

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: Instant domain-specific corpora to support human translators. In *Proceedings, 11th Annual Conference of the European Association for Machine Translation Conference*, Oslo, Norway, 247–252. Retrieved November 8, 2010, from http://www.mt-archive.info/EAMT-2006-Baroni.pdf

Bernardini, S. (1997). A 'trainee' translator's perspective on corpora. In *Proceedings, 1st International Conference on Corpus Use and Learning to Translate*, Bertinoro, Italy. Retrieved August 27, 2009, from http://web.archive.org/web/2003123890215/http://www.sslmit.unibo.it/cultpaps/paps.htm

Boulton, A. (2008). Looking for empirical evidence of DDL at lower levels. In B. Lewandowska-Tomaszczyk (Ed.), *Corpus linguistics, computer tools, and applications: State of the art* (pp. 581–598). Frankfurt, Germany: Peter Lang. Retrieved November 8, 2010, from http://hal.archives-ouvertes.fr/docs/00/38/49/08/PDF/2008_boulton_PALC_looking.pdf

Boulton, A. (in press). Bringing corpora to the masses: Free and easy tools for language learning. In N. Kübler (Ed.), *Selected papers from Teaching and Language Corpora 2006*. Amsterdam: Rodopi, Retrieved November 8, 2010, from http://hal.archives-ouvertes.fr/docs/00/32/69/80/PDF/XXXX_boulton_TaLC_interdisciplinary.pdf (7)

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. *ReCALL*, *17*(1), 47–64.

Castagnoli, S. (2006). Using the Web as a source of LSP corpora in the terminology classroom. In M. Baroni & S. Bernardini (Eds.), *Wacky! Working papers on the Web as corpus* (pp. 159–172). Bologna: Gedit. Retrieved August 27, 2009, from http://wackybook.sslmit.unibo.it/pdfs/castagnoli.pdf

Chambers, A. (2005). Integrating corpus consultation in language studies. *Language Learning & Technology*, *9*(2), 111–125. Retrieved August 27, 2009, from http://llt.msu.edu/vol9num2/chambers/

Cheng, H., & Dörnyei, Z. (2007). The use of motivational strategies in language instruction: The case of EFL teaching in Taiwan. *Innovation in Language Learning and Teaching*, *1*(1), 153–174. Retrieved November 8, 2010, from http://www.nottingham.ac.uk/english/research/cral/lib/exe/fetch.php?id=people%3Azoltan&;cache=cache&media=people:zoltan:2007_dornyei_cheng_illt.pdf.

Cheng, W., Warren, M., & Xu, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. *System*, *31*(2), 173–186.

Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the WAC4 Workshop at LREC 2008*, Marrakech, Morocco. Retrieved August 27, 2009, from http://clic.cimec.unitn.it/marco/publications/lrec2008/lrec08-ukWaC.pdf

Hadley, G. (2002). Sensing the winds of change: An introduction to data-driven learning. *RELC Journal*, *33*(2), 99–124. Retrieved November 8, 2010, from http://www.nuis.ac.jp/~hadley/publication/windofchange/windsofchange.htm

Ho, M. (1998). Culture studies and motivation in foreign and second language learning in Taiwan. *Language, Culture and Curriculum*, *11*(2), 165–182. Retrieved August 27, 2009, from http://pdfserve.informaworld.com/384088_731194254_907961081.pdf

Jackson, H. (1997). Corpus and concordance: Finding out about style. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 224–239). London: Longman.

Johns, T.F. (1991). Should you be persuaded: Two examples of data-driven learning. In T.F. Johns & P. King (Eds.), *Classroom concordancing* (pp. 1–13). Birmingham: ELR.

Johns, T.F. (1997). Contexts: The background, development and trialling of a concordance-based CALL program. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 100–115). London: Longman.

Kennedy, C., & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology*, *5*(3), 77–90. Retrieved August 28, 2009, from http://llt.msu.edu/vol5num3/kennedy/default.html

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the 11th EURALEX International Congress*, Barcelona, Catalonia. Retrieved November 8, 2010, from http://www.kilgarriff.co.uk/Publications/2008-KilgEtAl-euralex-gdex.doc

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress*, Lorient, France. Retrieved November 8, 2010, from http://kilgarriff.co.uk/Publications/2004-KilgRychlySmrzTugwell-SkEEuralex.rtf

Lai, H. (2008). *Learning English as an international language or not? A study of Taiwanese students' motivation and perceptions* (Unpublished PhD thesis). University of Warwick, UK. Retrieved November 8, 2010, from http://wrap.warwick.ac.uk/1111/1/WRAP_THESIS_Lai_2008.pdf

Maia, B. (1997). Making corpora: A learning process. In S. Bernardini & F. Zanettin (Eds.), *I corpora nella didattica della traduzione* (pp. 47–60). Bologna, Italy: CLUEB. Retrieved August 27, 2009, from http://www.sslmit.unibo.it/cultpaps/paps.htm

Savignon, S.J., & Wang, C. (2003). Communicative language teaching in EFL contexts: Learner attitudes and perceptions. *International Review of Applied Linguistics*, *41*, 223–249.

Seidlhofer, B. (2002). Pedagogy and local learner corpora: Working with learner-driven data. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 213–234). Amsterdam: John Benjamins.

Smith, S., Sommers, S., & Kilgarriff, A. (2008). Learning words right with the Sketch Engine: Meaningful lexical acquisition from corpora and the web. In *CamTESOL conference on English language teaching: Selected papers 4* (pp. 28–42). Phnom Penh, Cambodia. Retrieved November 8, 2010, from http://www.camtesol.org/Selected_Papers_Vol. 4_2008.pdf

Sun, Y.-C., & Wang, L.-Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16*(1), 83–94.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H.G. Widdowson* (pp. 125–144). Oxford, UK: Oxford University Press.

Thomas, J. (2008). Impatience is a virtue: Students and teachers interact with corpus data - now. In A. Frankenberg-Garcia (Ed.), *Proceedings of the 8th Teaching and Language Corpora Conference* (pp. 463–469). Lisbon: ISLA.

Tribble, C. (1997). Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. In J. Melia & B. Lewandowska-Tomaszczyk (Eds.), *PALC '97 Proceedings* (pp. 132–147). Lodz, Poland: Lodz University Press.

Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes*, *17*, 267–280.

Tyne, H. (2009). Corpus oraux par et pour l'apprenant [Spoken corpora by and for the learner]. In A. Boulton (Ed.), *Des documents authentiques oraux aux corpus: Questions d'apprentissage en didactique des langues* (pp. 91–111). Nancy, France: Mélanges CRAPEL. Retrieved November 8, 2010, from http://revues.univ-nancy2.fr/melangesCrapel/IMG/pdf/05_Tyne.pdf

Turnbull, J., & Burston, J. (1998). Towards independent concordance work for students: Lessons from a case study. *ON-CALL*, *12*(2), pp. 10–21. Retrieved August 27, 2009, from http://web.archive.org/web/20060505111607/http://www.cltr.uq.edu.au/oncall/turnbull122.html

Widdowson, H.G. (1978). *Teaching language as communication*. Oxford, UK: Oxford University Press.

Appendix 1. Instructions to students for WebBootCat project

This project looks scary because there are many tasks. In fact, most of these tasks are just to remind you how to build a corpus! Hopefully, you will end up with a corpus which is truly related to your major, which shows you relevant vocabulary in use.

Build a corpus, as we have been doing. Do points 1–14, and then answer the questions at the end. Write 300–400 words for C, maybe less or more for A and B.

(1) Choose 10 or more specialist words or multiword terms from your major to make a corpus
   (a) Perhaps your textbook has a glossary?
   (b) To save time later, use a seed term text (Notepad) file
(2) Give the corpus an appropriate name.
(3) Request a tagged corpus (this will take longer, but you need it to get Word Sketches).
(4) Don't change the number of URLs at first (because you are using more seed words than before).

(5) When the corpus is ready, extract the keywords, including multiword expressions
(6) Check the boxes of useful words
(7) Make a second corpus, with a new name, and get a keyword list.
(8) Note how many relevant words and phrases are on the list?
Now start again, and do some experiments, to see if you can make your corpus better. You don't have to try **everything**; if you do, you will drive yourself crazy!
(9) Uncheck some of the URLs at the first stage, so that some websites are ignored.
   (a) Or, you can do the same thing when the corpus is finished, by creating a sub-corpus based on certain websites (URLs) only.
(10) Use the cc (creative commons) option.
(11) Try changing some of the advanced options: number of URLs, number of tuples, page size.
You should now have one or two corpora that are quite good (=quite representative of your subject).
(12) Make Word Sketches for
   (a) The name of your subject (e.g. Economics).
   (b) Your original seed words.
   (c) Words which have a different meaning in your subject to the normal meaning (e.g. ''sheet'' in music, ''pie'' in statistics, ''death'' inlinguistics).
   (d) An ordinary word, with no special meaning (like ''computer'' or ''boy'').
(13) Click on some top collocations (just like regular Sketch Engine homework) and get concordances.
(14) OR (instead of 12 and 13) make short **sample** concordances (using beta and GDEX) for 12a,b,c,d)
(15) Do 12 and 13 (or 14) for an ordinary corpus, like BNC or UKWaC.

Questions

(A) To make a corpus, WBC uses Yahoo! It does an Internet search for groups of words (usually three words) called tuples. Try to describe in more detail how the program makes a corpus.
(B) To make a corpus, is it better to just use the default WBC options (for example, 10 URLs per query)? Or do you recommend using different options? Explain how you decided, giving examples.
(C) Using screen shots and other kinds of examples to help you, describe how well your corpus represents the topic you chose.

Appendix 2. Comments in students' final projects which address the research questions.

Student	Enjoyable/useful?	Learn English?	Ownership/corpus specialization?	Use after course?	Other skills?
X1	"Multi-words terms is also part I like."	Especially Economics terms. "I can take to expand my vocabulary."	"I chose 10 seed words from my textbook and some lecture from my economy professor."	"I will continue to use."	Gained knowledge from web relevant to my life
J1	"I want to choose WebBootCaT because I think it is more interesting than the other courses."		"I can also build a corpus about my major, and maybe I can use it to help my studies."	*"this may be my last experience to do this kind of homework"*	"Although it took me a long time to do it, I thought I got an experience that I have never had. Wow, I domesticated the WebBootCaT!!!!"
E1	His earlier corpora were "strange"; "it's a good chance to let me try again"				
T1	"I'm interested in making a corpus."		"We can have our own corpus by using WBC."		
D1	"Although it needs time to realize how to use, it is really interesting."		"I can use this chance to realize many functions of options and make the best corpus."	"When I have time in the summer vacation ... "	
N1		"During the time in studying history, I got some special words in this subject, some of them are normal to see, some of them are unique but representative."			

(*continued*)

Appendix 2. (*Continued*)

Student	Enjoyable/useful?	Learn English?	Ownership/corpus specialization?	Use after course?	Other skills?
S1		"That can make you know words easily because words are about your own interests."	"I find it is special to have your own corpus. It is unique!"		"I know more about my major after making a new corpus."
I1	"If I build my corpus, [rather than using a pubic corpus]I can search the words more efficiently."		"I build the corpuses, which means I make my own English information categories. It is personal, uncommon, and for what I need."		
V1					"I have learned a lot from this project, due to the tough process."
J2				*"I think I will have less opportunities to use the WebBootCat and Sketch Engine systems."*	
S2		"I think making a corpus is a good way to help us learning English."			
L1	*"I found that the results it showed are quite few and was much worse than ukWaC.That made me feel disappointed."*				

(*continued*)

Appendix 2. (*Continued*)

Student	Enjoyable/useful?	Learn English?	Ownership/corpus specialization?	Use after course?	Other skills?
P1	"And I think it is useful. Because when you build your own corpus, you can find what you want in a short time … When I built my corpus I think this is fun, and feel interesting."	"I think WBC is successful, because it really helps me to learn more English … "			
S3		"when I do words research with my corpus, I found very useful Collocations. The Concordances are quite similar to real life cases."	"I think my corpus is more useful in sketch about Accounting than general corpus such as UkWaC."		
R1			"Creating a specialized corpus could be useful when it comes to researching a particular subject or learning a subject in English. It is useful because of the different results which are much more relevant than searching on a much more general English corpus."		

(*continued*)

Appendix 2. (*Continued*)

Student	Enjoyable/useful?	Learn English?	Ownership/corpus specialization?	Use after course?	Other skills?
W1		"It is a practical tool for us to learn English."	"I also grasp how to build a corpus that represent to my topic."		"We are not only learning the language, we also learn the culture." "I have got further knowledge about data base system."
L2		"I think it's a good way to learn English in this way because it's useless to memorize the vocabulary without knowing how to use it."	"By creating your own corpus, you can develop your own professional area in languages supporting by WBC."	"I think the sketch engine and WEB will become my good partners of learning English."	

Note: Seventeen students made relevant comments. Negative comments are given *in italics*.